教程 7 : 贝叶斯

提交日期：2020.7.25

提交人：詹紫琦

目录

[1. 题目 3](#_Toc4338)

[1.1 贝叶斯 3](#_Toc17818)

[1.2 运行环境 3](#_Toc7337)

[2. 算法阐述或实验步骤说明 3](#_Toc17297)

[2.1变量说明 3](#_Toc26890)

[2.2 贝叶斯分类阐述及步骤说明 3](#_Toc3771)

[3. 实验结果与截图 4](#_Toc27974)

[3.1 贝叶斯离散值概率分类实验结果 4](#_Toc27904)

[3.2 贝叶斯连续值概率分类实验结果 6](#_Toc15668)

[4. 总结 6](#_Toc14878)

[4.1 总结 6](#_Toc10174)

[5. 参考文献 6](#_Toc27384)

# 题目

## 1.1 贝叶斯

贝叶斯[1]是一种统计学习方法，基本方法是在统计数据的基础上，根据条件概率公式，计算当前特征样本属于某个分类的概率，选择最大的概率分类项。为了简化运算，朴素贝叶斯分类算法假定任意属性对类别的影响与其他属性对类别的影响无关，即假设各个特征向量之间线性无关，这种假定称为类条件独立朴素假定，同时与逻辑回归，及其他分类算法不同的是，贝叶斯分类是一种判别模型，而不是一种生成模型，判别模型是根据已有数据集而进行判别分类。生成模型是生成一种逻辑关系同时满足输入输出。在本次贝叶斯实验中，既讨论了离散特征属性的贝叶斯分类，又讨论了连续型属性的贝叶斯分类。

## 1.2 运行环境

系统：Ubuntu20.04，python3.7，Anaconda集成工具Jupyter编写。

# 算法阐述或实验步骤说明

## 2.1变量说明

，a为x的特征属性。

类别集合。

条件概率公式。

满足y类别的集合。

属性种类数。

## 2.2 贝叶斯分类阐述及步骤说明

首先读取数据集，得到，得到满足条件的每一个特征属性，然后分别计算，找出中最大的概率值，并将y作为x的所属类别，大概步骤如上，但是如果一个个计算，再进行比较会耗费许多时间，这里使用贝叶斯公式如公式2-1：

 （2-1）

一般来说，公式2-1的分母为常数，所以比较大小只需要比较分子，而分子的构成又能使用公式2-2得到：

 （2-2）

其中可由公式2-3得到：

 （2-3）

这里计算概率的同时使用了拉普拉斯平滑，为了避免在概率连乘时产生乘零的结果。所以使用了拉普拉斯平滑。可由公式2-4得到：

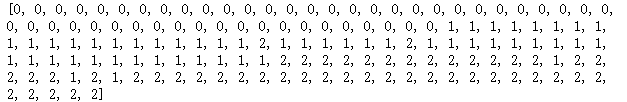
 （2-4）

同样公式2-4也使用了拉普拉斯平滑。计算出先验概率后，进行比较，最后判定x所属的类别。

# 实验结果与截图

## 3.1 贝叶斯离散值概率分类实验结果

首先读取数据集，这里用的是鸢尾花数据集，初步分析数据类别，分为三类，分别用0，1，2替代。得到的数据图像为3-1：这里我们首先使用离散概率计算来进行贝叶斯分类，但由于鸢尾花数据的各个特征值都为连续型数据，这里我们设定阈值，分别将花萼长度，花萼宽度，花瓣长度，花瓣宽度设置两个阈值，这样每一个特征都能被分为三个类别，我们分别用1，2，3替代。其中花萼长度的阈值为5，6；花萼宽度的阈值为3，3.5；花瓣长度的阈值为3，5；花瓣宽度的阈值为0.8，1.6。设置阈值后得到新的离散型数据，初步展示前三列如图像3-2。得到处理后的数据集后进行贝叶斯概率计算，对每一项数据进行分类，并统计分类正确率，的得到的贝叶斯离散分类的正确率为96.67%





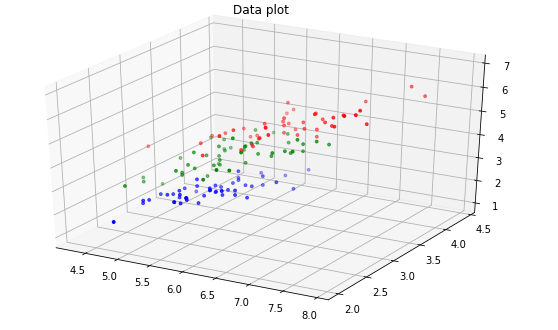


图3-1 鸢尾花数据集前三维数据不同类别图像点云

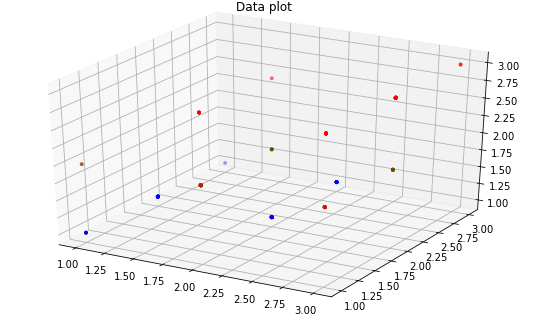


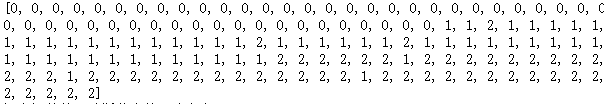
图3-2 通过阈值分类后的离散前三维数据不同类别图像点云

## 3.2 贝叶斯连续值概率分类实验结果

将鸢尾花数据通过不同的阈值进行分类后计算出的分类正确率为99.67%，然后改变策略，使用鸢尾花原始数据，通过公式3-1进行概率计算：

 （3-1）

其中和分别是第c类样本在第i个属性上取值的均值和标准差。最后跟离散值同样的评价标准，计算分类正确率，得到最后的分类正确率为96.00%。





# 总结

## 4.1 总结

贝叶斯分类使用的是极大似然估计，通过假设各特征线性无关，从而通过概率连乘的方式得到分类的概率，最终找出最大可能性的分类类别。估计类条件概率的一种常用策略是先假定其具有某种确定的概率分布形式，再基于训练样本对概率分布的参数进行估计。事实上，概率模型的训练过程就是参数估计过程，对于参数估计，统计学界的两个学派分别提供了不同的解决方案，一个是优化似然函数等准则来确定参数值，另一种是认为参数是未观察到的随机变量，也可能有分布，因此，假定参数服从一个先验分布，然后基于观测到的数据来计算参数的后验分布。

# 参考文献

[1] 周志华等.机器学习（西瓜书）